PCT/JP00/01177

15.03.00

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

JP00/11144

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９９年 ３月 １日

出願番号
Application Number: 平成１１年特許願第０５２１５６号

出願人
Applicant (s): 松下電器産業株式会社

２０００年 １月１４日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

出証番号 出証特平１１－３０９３６００

【書類名】　特許願

【整理番号】　2036610001

【提出日】　平成11年 3月 1日

【あて先】　特許庁長官

【国際特許分類】　G06K 09/20

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　物部　祐亮

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　広瀬　篤嗣

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　梅林　明人

【特許出願人】

【識別番号】　000005821

【氏名又は名称】　松下電器産業株式会社

【代理人】

【識別番号】　100083172

【弁理士】

【氏名又は名称】　福井　豊明

【手数料の表示】

【予納台帳番号】　009483

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9713946

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書画像処理装置及びそのタイトル領域抽出方法

【特許請求の範囲】

【請求項１】　文書画像を複数の領域に分割する領域分割手段と、該領域分割手段によって分割された各領域について領域平均文字サイズを算出した後、該領域平均文字サイズに基づいて全領域の中からタイトル領域を抽出するタイトル領域抽出手段とを備えた文書画像処理装置において、

全領域の文字の平均高さに相当する全平均文字サイズを算出した後、該全平均文字サイズに抽出パラメータを乗算した抽出判定値と上記領域平均文字サイズとを比較し、上記抽出判定値より大きい領域平均文字サイズの領域をタイトル領域として抽出する上記タイトル領域抽出手段を備えたことを特徴とする文書画像処理装置。

【請求項２】　上記タイトル領域抽出手段が、複数段階の抽出パラメータを用いて複数段階の上記抽出判定値を算出する請求項１に記載の文書画像処理装置。

【請求項３】　上記タイトル領域抽出手段が、複数段階の抽出パラメータを用いて複数段階の上記抽出判定値を算出するとともに、抽出した段階を示すレベル属性を対応付けてタイトル領域を抽出する請求項１に記載の文書画像処理装置。

【請求項４】　上記タイトル領域抽出手段が、領域平均文字サイズの最大値を全平均文字サイズで除算した値に基づいて上記複数段階の抽出パラメータを決定する請求項２または３に記載の文書画像処理装置。

【請求項５】　上記タイトル領域抽出手段が、全平均文字サイズおよび領域平均文字サイズを、所定割合より大きい文字および所定割合より小さい文字を除外した文字より算出するトリム平均を用いる請求項１に記載の文書画像処理装置。

【請求項６】　文書画像を複数の領域に分割し、各領域について文字の平均高さに相当する領域平均文字サイズを算出した後、該領域平均文字サイズに基づいて全領域の中からタイトル領域を抽出する文書画像処理装置のタイトル領域抽

出方法において、

全領域の全平均文字サイズを算出した後、該全平均文字サイズに抽出パラメータを乗算した抽出判定値と上記領域平均文字サイズとを比較し、上記抽出判定値より大きい領域平均文字サイズの領域をタイトル領域として抽出することを特徴とする文書画像処理装置のタイトル領域抽出方法。

【請求項７】　複数段階の抽出パラメータを用いて複数段階の上記抽出判定値を算出する請求項６に記載の文書画像処理装置のタイトル領域抽出方法。

【請求項８】　複数段階の抽出パラメータを用いて複数段階の上記抽出判定値を算出するとともに、抽出した段階を示すレベル属性を対応付けてタイトル領域を抽出する請求項６に記載の文書画像処理装置のタイトル領域抽出方法。

【請求項９】　領域平均文字サイズの最大値を全平均文字サイズで除算した値に基づいて上記複数段階の抽出パラメータを決定する請求項７または８に記載の文書画像処理装置のタイトル領域抽出方法。

【請求項１０】　所定割合より大きい文字および所定割合より小さい文字を除外した文字の平均値を算出するトリム平均を用いて全平均文字サイズおよび領域平均文字サイズを算出する請求項６に記載の文書画像処理装置のタイトル領域抽出方法。

【請求項１１】　上記文書画像が複数頁の文書画像である請求項６～１０に記載の文書画像処理装置のタイトル領域抽出方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書画像を画像データとして記憶・管理する文書画像処理装置と文書画像処理方法に関し、特に、上記文書画像からタイトル領域を抽出する上記装置と方法に関するものである。

【０００２】

【従来の技術】

データ記憶装置の容量が著しく増加してきたことに伴って、スキャナ等から読み込んだ紙文書を画像データである文書画像として記憶・管理する文書画像処理

装置が急速に普及してきている。

【０００３】

このような文書画像処理装置では、データ記憶装置に記憶された複数の文書画像の中から所望の文書画像を検索できるようにするため各文書画像にキーワードとなる文字列を対応付けて登録するようにしており、このキーワードは、文書画像を記憶する際にユーザが手入力していた。しかしながら、このようなキーワード入力作業をユーザが行うことは、文書数が多くなると作業量が膨大になるため現実的でない。そこで、近年では、文書画像に対して文字認識を行い、この認識結果である文字列をキーワードとすることによって、人手を介さずにキーワードを付与できるようにした装置も出現している。

【０００４】

ところで、文書画像を効率良く検索するためには、適切なキーワードを付与しておくことが重要である。すなわち、ディスプレイに一覧表示された複数キーワードの中から所望の文書画像に対応するキーワードを特定するのが一般的な検索形態であるが、このようなキーワードを迅速に特定するためには、各キーワードが文書の内容を端的に表した内容でなければならない。

【０００５】

特開平８－２０２８５９号公報では、タイトル文字列が属する領域（以下「タイトル領域」という。）を文書画像から抽出した後、このタイトル領域画像に対して文字認識を行い、この認識結果であるタイトル文字列をキーワードとする方法を提案している。タイトル文字列は文書の内容を端的に表した内容であるため、このようなタイトル領域抽出方法を採用した文書画像処理装置によれば所望の文書画像に対応するキーワードを迅速に特定できる。

【０００６】

【発明が解決しようとする課題】

上記従来のタイトル領域抽出方法では、タイトル文字が当該文書画像に属する全ての文字の中で最も大きいサイズであるという観点から、文書画像を複数の領域（隣接する文字矩形相互を統合した領域）に分割して各領域内の平均文字サイズを算出し、この平均文字サイズが最も大きい領域をタイトル領域として抽出す

るようにしている。従って、このようなタイトル領域抽出方法によって抽出されるタイトル領域の数は、１つの文書画像につき当然１つとなる。

【０００７】

しかしながら、近似した内容の複数の文書が存在する場合、タイトルも近似した内容となるのが通常であるため、上記従来のタイトル領域抽出方法には、近似した内容の文書が多数存在する場合所望の文書画像に対応するキーワードを迅速に特定できないという問題があった。

【０００８】

上記問題を回避するため紙文書を作成する段階で似た内容のタイトルを付けないようにしてもよいが、このような準備作業をユーザに要求することは好ましくない。

【０００９】

本発明は上記従来の事情に基づいて提案されたものであって、１つの文書画像から複数のタイトル領域を抽出できるようにした文書画像処理装置及びそのタイトル領域抽出方法を提供することを目的とするものである。

【００１０】

【課題を解決するための手段】

本発明は上記目的を達成するために以下の手段を採用している。すなわち、図１に示すように、文書画像を複数の領域に分割する領域分割手段３と、該領域分割手段３によって分割された各領域について領域平均文字サイズを算出した後、該領域平均文字サイズに基づいて全領域の中からタイトル領域を抽出するタイトル領域抽出手段４とを備えた文書画像処理装置を前提としている。

【００１１】

ここで、上記タイトル領域抽出手段４は、全領域の文字の平均高さに相当する全平均文字サイズを算出した後、該全平均文字サイズに抽出パラメータを乗算した抽出判定値と上記領域平均文字サイズとを比較し、上記抽出判定値より大きい領域平均文字サイズの領域をタイトル領域として抽出するようにしている。このようにすれば、上記抽出判定値より大きい領域平均文字サイズの領域であればタイトル領域として抽出されるため、１つの文書画像から複数のタイトル領域を抽

出できることになる。

【００１２】

また、上記タイトル領域抽出手段４が、複数段階の抽出パラメータを用いて複数段階の上記抽出判定値を算出するようにしてもよい。このようにすれば、複数段階の抽出判定値に基づいて抽出判定がなされることになるため、タイトル領域だけでなくサブタイトル領域（タイトル文字より若干小さなサイズの文字からなるサブタイトル文字列が属する領域）をも抽出できる。

【００１３】

更に、上記タイトル領域抽出手段４が、領域平均文字サイズの最大値を全平均文字サイズで除算した値に基づいて上記複数段階の抽出パラメータを決定するようにしてもよい。抽出パラメータを固定値とするのではなく領域平均文字サイズの最大値等に基づいて算出した方が良好な抽出判定値が得られる。

【００１４】

上記全平均文字サイズ、領域平均文字サイズを求めるについて、所定割合より大きい文字および所定割合より小さい文字を除外したトリム平均を用いるとにより精度を上げることができる。

【００１５】

【発明の実施の形態】

以下に本発明の実施の形態を図面に従って詳細に説明する。

（第１の実施の形態）

図１は、本発明を適用した文書画像処理装置の概略機能ブロック図であり、以下、その構成を文書画像登録手順とともに説明する。

【００１６】

まず、例えばスキャナ等の文書画像入力手段１が紙文書を光電変換して多値画像データである文書画像８ａを得、該文書画像は画像処理手段１１ａで記憶に適した処理（例えば圧縮処理）がなされて記憶手段８の文書画像エリアＡａに登録される。もちろん画像処理手段１１ａを設けないで多値画像データのまま文書画像エリアＡａに登録しておいてもよい。

【００１７】

上記文書画像入力手段１よりの文書画像は上記画像処理手段１１ａに入力されるとともに、画像処理手段１１ｂにも入力されここで２値画像データに変換されて画像メモリ７に格納される。このように画像メモリ７に文書画像が格納された状態で、文字矩形生成手段２は上記画像メモリ７に記憶された文書画像を参照して、以下のラベリング処理を行う。このラベリング処理とは、注目する黒画素（以下「注目画素」という。）の上、右上、右、右下、下、左下、左、左上の８方向に隣接する画素のうち黒画素について当該注目画素と同一のラベル値（識別情報）を与える処理である。すなわち、図７に示すようにＷ１・Ｗ２・Ｗ３・Ｗ４・Ｗ６・Ｗ７・Ｗ８・Ｗ９の８画素が注目画素Ｗ５に連結する場合、文字矩形生成手段２は、黒画素であるＷ２・Ｗ３・Ｗ８に注目画素Ｗ５と同一のラベル値を与える。このようなラベリング処理を行うことによって、文書画像内の黒画素連結成分（連続する黒画素）毎に同一ラベル値を与えることができる。

【００１８】

次いで、文字矩形生成手段２は、上記のように同一ラベル値を与えた黒画素連結成分を切り出すことによって文字矩形を生成し、この文字矩形を領域分割手段３に渡す。ここで、文字矩形とは黒画素連結成分の外接矩形を意味する。尚、文字によっては１つの黒画素連結成分で構成されていない場合もあり、このことを考慮して、上記ラベリング処理を行う前に文書画像中の黒画素領域を膨張させる処理をしておくこともできる。すなわち、注目する黒画素に隣接する８個の画素を黒画素に変換するという処理であり、この処理を適切な回数（通常２、３回）だけ施すことにより、黒画素の領域が拡大され１つの文字内で分離していた黒画素連結成分を１つに結合することができる。このような処理を行った上で、上記ラベリング処理を行うことにより、上記文字矩形を正しく文字毎に生成することが可能となる。

【００１９】

上記文字矩形生成手段２の処理が終わると領域分割手段３は、各文字矩形について近傍を調べ、相互に隣接する文字矩形を統合することによって文書画像の領域を分割する。例えば図８に示す文字矩形Ｃ１～Ｃ１２を受けた領域分割手段３は、文字矩形Ｃ１～Ｃ４・Ｃ５～Ｃ９・Ｃ１０～Ｃ１２をそれぞれ統合すること

によって文書画像を領域Ｅ１・Ｅ２・Ｅ３に分割する。このような領域分割処理を行うことによって、文書画像の領域を文字列毎に分割することができる。なお、文字矩形が相互に隣接している状態であるのか、あるいは、行間であるのか等の区別は文字間、行間に関する適当な閾値を用いて判定するようにしている。

【００２０】

以上の結果、文書画像内における全ての文字サイズ（後述する）・分割された領域数・各領域内の文字矩形の数などの情報が得られる。本発明では、分割された各領域に対して１から始まる通し番号を付すとともに各領域に属する文字矩形に対しても１から始まる通し番号を付すようにしており、以下、ｎ番目の領域内の文字矩形数を$\mathrm{Num\ Char}_n$、ｎ番目の領域内におけるｍ番目の文字サイズを$\mathrm{Size\ Char}_{n,m}$と表す。

【００２１】

ところで図９に示すように、文字矩形の幅Ｗ１～Ｗ４および面積Ａ１～Ａ４は同一ポイント数の文字フォントを使用している場合であっても文字の種類に依存して大きく変動するのに対して、文字矩形の高さＨ１～Ｈ４はこのような変動が小さい。従って本発明では、文字フォントのポイント数が比較的正確に反映される“文字矩形の高さ”を上記文字サイズとして採用するようにしている。

【００２２】

ここで、タイトル領域抽出手段４は、上記のように分割された全領域のうち所定の領域のみをタイトル領域として抽出する。以下、このタイトル領域抽出処理を図２に示すフローチャートに従って説明する。

【００２３】

まず、タイトル領域抽出手段４は各領域について領域平均文字サイズを算出する（図２、ステップ１）。この領域平均文字サイズとは１領域に属する全ての文字サイズの平均値であり、ｎ番目の領域における領域平均文字サイズ$\mathrm{SizeReg}_n$は、当該領域に属する全ての文字サイズ$\mathrm{Size\ Char}_{n,m}$の加算値を当該領域内の文字数$\mathrm{Num\ Char}_n$で除算した値となる。この関係を次式に示す。

【００２４】

【数１】

$SizeReg_n = \Sigma SizeChar_{n,m} / NumChar_n$

次いで、上記のように算出した各領域の領域平均文字サイズ$SizeReg_n$と領域内の文字数$Num\ Char_n$とから、文書画像内の全平均文字サイズSizeAll を次式によって算出する（図２、ステップ２）。

【００２５】

【数２】

$SizeALL = \Sigma (SizeReg_n \times NumChar_n) / \Sigma NumChar_n$

なお、領域平均文字サイズ$SizeReg_n$および全平均文字サイズSizeAll の算出方法は上記した方法に限定されるものではなく、例えば、後に説明するトリム平均（最小値側および最大値側から所定割合例えば１０％のデータを除外したうえで平均値を算出する方法）を採用することもできる。

【００２６】

ここで、タイトル領域抽出手段４は、以下に示す抽出判定式が成立するか否かに基づいてタイトル領域の抽出判定を行う。

【００２７】

【数３】

$SizeReg_n >= SizeAll \times \alpha$

すなわち、上記のように算出した全平均文字サイズSizeAll に抽出パラメータαを乗算した値（抽出判定値）と各領域の領域平均文字サイズ$SizeReg_n$とを比較し、この抽出判定式が成立する領域のみをタイトル領域として抽出する（図２、ステップ３→４→５）。なお、抽出パラメータαは1.0 より大きな定数とし、1.2 程度の値とするのが好ましい。

【００２８】

以上の手順を繰り返し全ての領域について抽出判定が行われると（図２、ステップ３で“ＮＯ”）、タイトル領域抽出処理を終了し、ここで抽出された各タイトル領域画像８ｂは記憶手段８のタイトルエリアＡｂに登録される。

【００２９】

次いで、文字認識手段５は、上記のように抽出されたタイトル領域画像を文書画像から切り出し、このタイトル領域画像に対して文字認識を行うことによって

文字コード列であるタイトル文字列を得た後、このタイトル文字列を文書登録手段６に渡す。

【００３０】

上記タイトル文字列を受けた文書登録手段６は、記憶手段８での文書画像８ａの格納ポインタ、上記タイトル領域画像８ｂの格納ポインタ・上記タイトル文字列・文書画像内におけるタイトル領域の位置およびサイズからなる登録情報を記憶手段８上のテーブルエリアＡｃに形成された登録情報管理テーブル８ｃ（図５参照）に登録する。ここで、上記文書画像８ａの格納ポインタは上記記憶手段８の文書画像エリアＡａより得られ、上記タイトル画像８ｂの格納ポインタは上記記憶手段８のタイトルエリアＡｂより得られ、更に、タイトル領域の位置とサイズは文字認識手段５より得られることになる。

【００３１】

このように登録情報管理テーブル８ｃが生成されると、以降に、キーボードやポインティングデバイス等からなる指示入力手段９より文書画像の検索が指示入力されると、表示制御手段１０は、上記のように記憶されたタイトル領域画像およびタイトル文字列を図示しないディスプレイにリスト表示する（図１０(II)参照）。

【００３２】

そして、上記リスト表示から所望のタイトル（タイトル領域画像またはタイトル文字列）が選択されたときには、このタイトルに対応する文書画像をディスプレイに表示する。このとき、図１０(II)に示すように、矩形枠Ｆで囲むなどして文書画像内におけるタイトル領域を明示するのが好ましい。このような矩形枠Ｆは、登録情報管理テーブル８ｃに登録されているタイトル領域の位置およびサイズに基づいて生成できる。

【００３３】

また、上記のようにディスプレイに表示されたリストからいずれか１つを選択する方法に加えて、指示入力手段９より特定のキーワードを入力し、該キーワードに該当するタイトルが登録情報管理テーブル８ｃに登録されているとき、対応する文書画像を表示するようにしてもよいことはもちろんである。

【００３４】

以上のように本実施の形態によれば、抽出判定値より大きい領域平均文字サイズの領域であればタイトル領域として抽出する構成としているため、１つの文書画像から複数のタイトル領域を抽出できる。従って、似た内容の文書が多数存在する場合であっても、所望の文書画像に対応するキーワード（タイトル）を迅速に特定できる。

【００３５】

なお、上記の説明では、タイトル領域抽出処理において抽出判定式の成立する領域が存在しなかった場合の手順については言及していないが、このような場合には、タイトル領域が抽出されなかった旨をディスプレイ表示するとともにキーワードとなる文字列を入力するようユーザに対して要求し、この要求に対してユーザが文字列を入力すると、この文字列を当該文書画像のタイトル文字列として用いるようにしている。

（第２の実施の形態）

上記第１の実施の形態では、抽出判定値より大きい領域平均文字サイズの領域であれば、領域平均文字サイズの大小を区別することなく同様にタイトル領域として抽出する構成としている。従って、タイトル文字より若干小さなサイズの文字からなるサブタイトル文字列はリスト表示せずタイトル文字列のみをリスト表示する処理など、領域平均文字サイズの大小に基づいた適切な処理を行うことができない。本実施の形態では、複数段階の抽出パラメータを用いて複数段階の抽出判定値を算出するとともにレベル属性（抽出した段階を示す情報）と対応付けてタイトル領域を抽出する構成とすることによって上記した問題を解消しており、以下、その構成を第１の実施の形態と異なる点のみ説明する。

【００３６】

上記第１の実施の形態と同様の手順で領域平均文字サイズ$SizeReg_n$および全平均文字サイズSizeAll を算出したタイトル領域抽出手段４は、以下に示す複数段階の抽出判定式が成立するか否かに基づいて複数段階の抽出判定を行う。

【００３７】

【数４】

$SizeReg_n >= SizeAll \times \alpha_p$

上式におけるα_pは、ｐ段階（レベルｐ）の抽出パラメータであり、〔数５〕の条件を満たすように値を設定しておく。例えば、５段階の抽出判定を行う場合には、α_1=1.5 、α_2=1.3 、α_3=1.2 、α_4=1.15、α_5=1.1 程度とするのが好ましい。

【００３８】

【数５】

$1.0 < \alpha_p < \cdots < \alpha_3 < \alpha_2 < \alpha_1$

図３に示すフローチャートを用いて説明すると、タイトル領域抽出手段４は、レベル１から順に全レベルの抽出判定を行い（図３、ステップ１４→１５→１４）、全レベルにおいて抽出判定式が成立しなかった場合には、この領域をタイトル領域として抽出せず、次の領域について抽出判定を行う（図３、ステップ１４→１３→１４→１５）。一方、いずれかのレベルにおいて抽出判定式が成立した場合には、この領域を当該レベルのタイトル領域として（上記レベル属性を対応付けて）抽出した後、次の領域について抽出判定を行う（図３、ステップ１５→１６→１３→１４→１５）。

【００３９】

以上の手順を繰り返し全ての領域について抽出判定が行われると（図３、ステップ１３で“ＮＯ”）、タイトル領域抽出処理を終了する。

【００４０】

なお、抽出判定式の成立する領域が存在しなかった場合ユーザが入力した文字列をタイトル文字列として用いる点は上記第１の実施の形態と同様であり、このタイトル文字列のレベル属性はレベル１、全レベル数も１としている。

【００４１】

図６は、本実施の形態における登録情報管理テーブル８ｃの説明図であり、上記第１の実施の形態において示した構成にレベル属性フィールドと全レベル数フィールドとを加えた構成としている。そして文書登録手段６は、例えば５段階の抽出判定においてレベル１で抽出された領域がある場合、この領域に対応する全レベル数フィールドには“５”を、レベル属性フィールドには“１”をそれぞれ

登録する。

【００４２】

図１１は、本実施の形態の検索時においてディスプレイに表示される内容を示す図であり、上段にリスト表示するタイトルのレベル属性を指示入力手段９より範囲指定できるようにしている。そして、表示制御手段１０は、登録情報管理テーブル８ｃのレベル属性フィールドと全レベル数フィールドとを参照することによって上記のように指定された範囲内のタイトルのみをディスプレイにリスト表示する。

【００４３】

以上のように本実施の形態によれば、複数段階の抽出パラメータを用いて複数段階の抽出判定値を算出するとともにレベル属性と対応付けてタイトル領域を抽出する構成としているため、サブタイトル文字列はリスト表示せずタイトル文字列のみをリスト表示する処理など領域平均文字サイズの大小に基づいて、異なる処理を行うことができる。

（第３の実施の形態）

上記第２の実施の形態では、複数段階の抽出パラメータを予め設定する（固定値とする）構成としているが、このような抽出パラメータは入力された文書画像の特性に応じて決定するのが好ましい。本実施の形態では、領域平均文字サイズの最大値を全平均文字サイズで除算した値に基づいて複数段階の抽出パラメータを決定する（図４、ステップ２３参照）ようにしており、以下、その構成を第２の実施の形態と異なる点のみ説明する。

【００４４】

上記第２の実施の形態と同様の手順で領域平均文字サイズ$SizeReg_n$ および全平均文字サイズSizeAll を算出したタイトル領域抽出手段４は、まず、領域平均文字サイズの最大値max $\{SizeReg_n\}$ を全平均文字サイズSizeAll で除算した値α_1 を次式によって算出する。

【００４５】

【数６】

$$\alpha_1 = \max \{SizeReg_n\} / SizeAll$$

次いで、タイトル領域抽出手段４は、上記のように算出したα_1と当該抽出判定の全レベル数P(P>=1)とから、各レベルの抽出パラメータα_pを次式によって決定する。

【００４６】

【数７】

$$\alpha_p = \alpha_1 - (p - 1) \times (\alpha_1 - 1) / P$$

例えばα_1が1.5 で５段階の抽出判定を行う場合、各レベルの抽出パラメータα_1～α_5は以下のようになる。

【００４７】

【数８】

$$\alpha_1 = 1.5 - (1 - 1) \times (1.5 - 1) / 5 = 1.5$$

$$\alpha_2 = 1.5 - (2 - 1) \times (1.5 - 1) / 5 = 1.4$$

$$\alpha_3 = 1.5 - (3 - 1) \times (1.5 - 1) / 5 = 1.3$$

$$\alpha_4 = 1.5 - (4 - 1) \times (1.5 - 1) / 5 = 1.2$$

$$\alpha_5 = 1.5 - (5 - 1) \times (1.5 - 1) / 5 = 1.1$$

このように【数7】によれば、上記のように算出したα_1から1.0の間で等間隔になるように各レベルの抽出パラメータα_pを決定することができる。

【００４８】

以降の手順は、上記のように決定した抽出パラメータを用いて抽出判定を行う点を除いて第２の実施の形態と同様であるため説明を省略する。

【００４９】

ただし上記した方法には、文書画像内にタイトル領域が存在しない場合、α_1が例えば1.03など1.0 付近の値となるため本文の領域をタイトル領域として誤抽出してしまうという不具合がある。そこで本発明では、例えば1.05など所定値以下となる抽出パラメータは採用しないようにしている。

【００５０】

また、各レベル間の抽出パラメータの差が例えば0.03など所定値以下となると、良好な抽出判定ができないため、上記抽出パラメータの差が上記所定値（0.03）となるように抽出パラメータの設定値を修正するようにしている。すなわち上

記の場合、α_1 から順に0.03ずつ減算した値を各レベルの抽出パラメータとして設定する。

【００５１】

以上の結果全レベル数Ｐが減少する場合もあるが、このような場合には、実際のレベル数（全レベル数Ｐから減少レベル数を減じた値）を全レベル数Ｐとして登録情報管理テーブル８ｃの全レベル数フィールドに設定する。

【００５２】

以上のように本実施の形態によれば、抽出パラメータを固定値とするのではなく、入力された文書画像の特性に応じて決定する構成としているため良好な抽出判定を行うことができる。

（第４の実施の形態）

上記の各実施の形態においては、全平均文字サイズの算出に比較的サイズの大きいタイトル領域の文字も算入され、また、サイズの小さいコンマ、ピリオド、句読点も算入されるので、精度が低くなる傾向がある。そこで、文書画像の全文字から、所定割合（例えば９０％）より大きいサイズの文字と、所定割合（例えば１０％）より小さいサイズの文字を除外した文字から全平均文字サイズを算出する、いわゆるトリム平均を利用する。更に、領域平均文字サイズを算出するときにも、同様の問題が発生するところから、領域平均文字サイズの算出についても上記トリム平均を用いることもできる。

【００５３】

これによって、全平均文字サイズ、および領域平均文字サイズとも、ピリオド、コンマ、句読点を除外した文字サイズを求めることができ、より精度の高い値が得られることになる。

【００５４】

ここで、上記各実施の形態では領域平均文字サイズより、全平均文字サイズを算出しているが、同じ方法をこのトリム平均を用いる場合に適用すると、領域毎にサイズの大きい文字と小さい文字を除外することになるため、全平均文字サイズの算出においてタイトル領域に含まれるすべての文字を除外することができない。従ってここでは、全平均文字サイズを算出するときに、あらためて文書画像

中の全文字を対象として処理を行っている。

【００５５】

但し、このトリム平均を用いる方式を使用するにしても、上記抽出パラメータとして、実施の形態１の所定値、あるいは実施の形態２、３の段階値のいずれを用いてもよいことはもちろんである。

【００５６】

なお、上記の各実施の形態の説明では、文書画像となる文書の枚数については言及していないが、紙文書の枚数は特に限定されるものではない。すなわち、１枚であっても複数枚であっても、各頁に同じ抽出パラメータを用いる限り同様の効果が得られる。特に、実施の形態２、３においては、複数頁に対して同じ抽出パラメータ用いることにより、論文データのように複数頁に渡る単一文書から、タイトル、サブタイトルを正しく抽出することができる。

【００５７】

また、上記の説明では、文字矩形の高さを文字サイズとして採用することとしているが、文字矩形の幅・面積を文字サイズとして採用してもよい。

【００５８】

尚、図１の説明において、記憶手段８の前段と画像メモリ７の前段に画像処理手段１１ａ、１１ｂを設けて、タイトル抽出用の文書画像は２値画像データを用い、記憶手段８の文書画像エリアＡａに登録される文書画像データとして、圧縮画像あるいは多値画像データを用いることができるようになっている。これによって、上記のように抽出されたタイトルに基づく検索処理の結果得られた文書画像をカラーで表示する等の多様な表示方法が可能となる。

【００５９】

【発明の効果】

以上のように本発明によれば、抽出判定値より大きい領域平均文字サイズの領域をタイトル領域として抽出するようにしているため、１つの文書画像から複数のタイトル領域を抽出できる。また、複数段階の抽出パラメータに基づいて複数段階の抽出判定をすることもでき、更に、この複数段階の抽出パラメータを入力された文書画像の特性に応じて決定できる。また、全平均文字サイズの算出、あ

るいは領域平均文字サイズの算出に、大きい方の所定割合と小さい方の所定割合に属する文字を除外して算出するトリム平均を用いると、より精度を上げることができる。

【図面の簡単な説明】

【図１】

本発明の文書画像処理装置の概略機能ブロック図である。

【図２】

第１の実施の形態におけるタイトル領域抽出処理のフローチャートである。

【図３】

第２の実施の形態におけるタイトル領域抽出処理のフローチャートである。

【図４】

第３の実施の形態におけるタイトル領域抽出処理のフローチャートである。

【図５】

第１の実施の形態における登録情報管理テーブルの説明図である。

【図６】

第２の実施の形態における登録情報管理テーブルの説明図である。

【図７】

ラベリング処理の説明図である。

【図８】

領域分割処理の説明図である。

【図９】

文字矩形の高さ・幅・面積の関係を示す図である。

【図１０】

第１の実施の形態の検索時においてディスプレイに表示される内容を示す図である。

【図１１】

第２の実施の形態の検索時においてディスプレイに表示される内容を示す図である。

【符号の説明】

１　文書画像入力手段

２　文字矩形生成手段

３　領域分割手段

４　タイトル領域抽出手段

５　文字認識手段

６　文書登録手段

７　画像メモリ

８　記憶手段

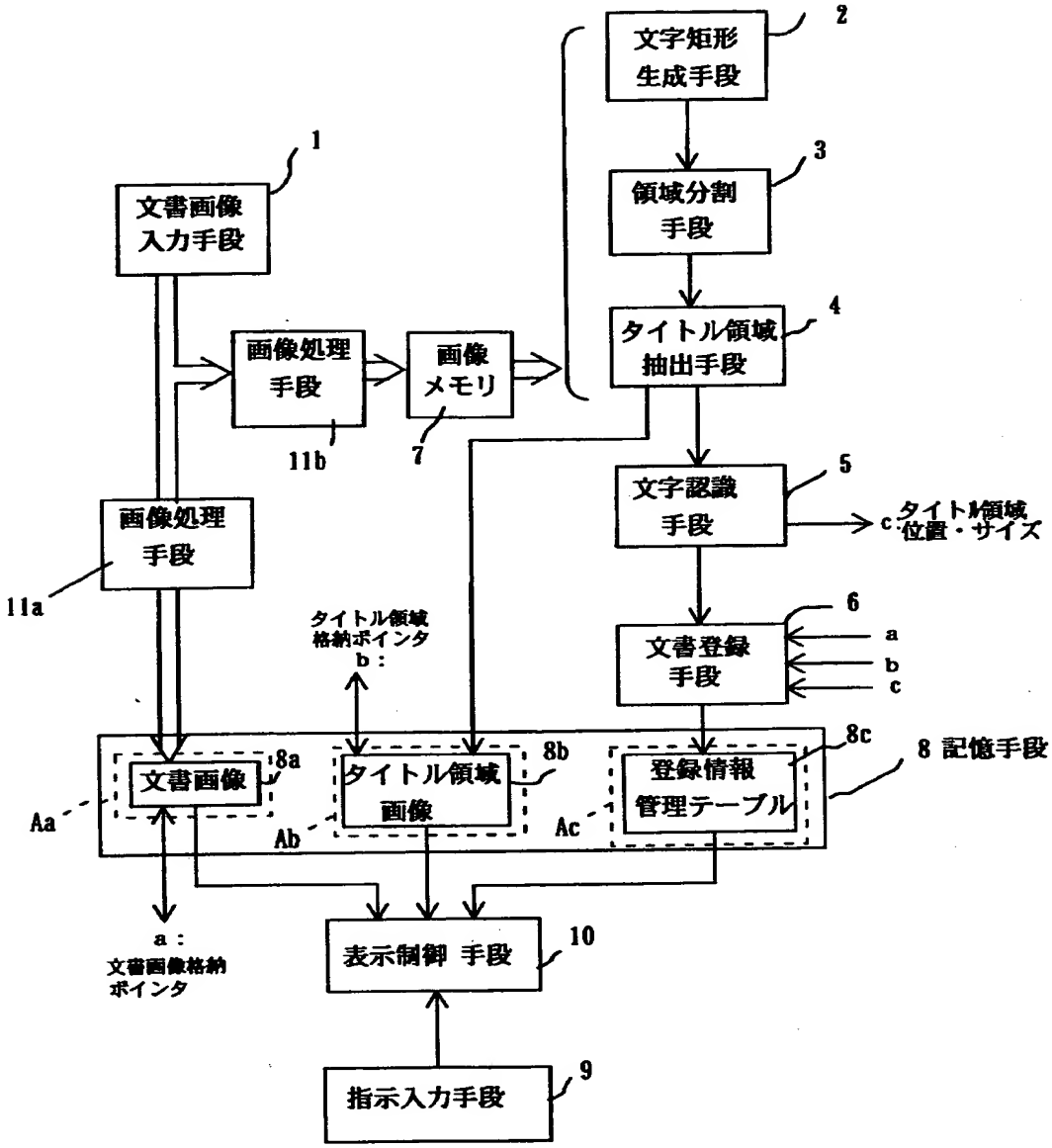
９　指示入力手段

１０　表示制御手段

１１（１１ａ、１１ｂ）画像処理手段

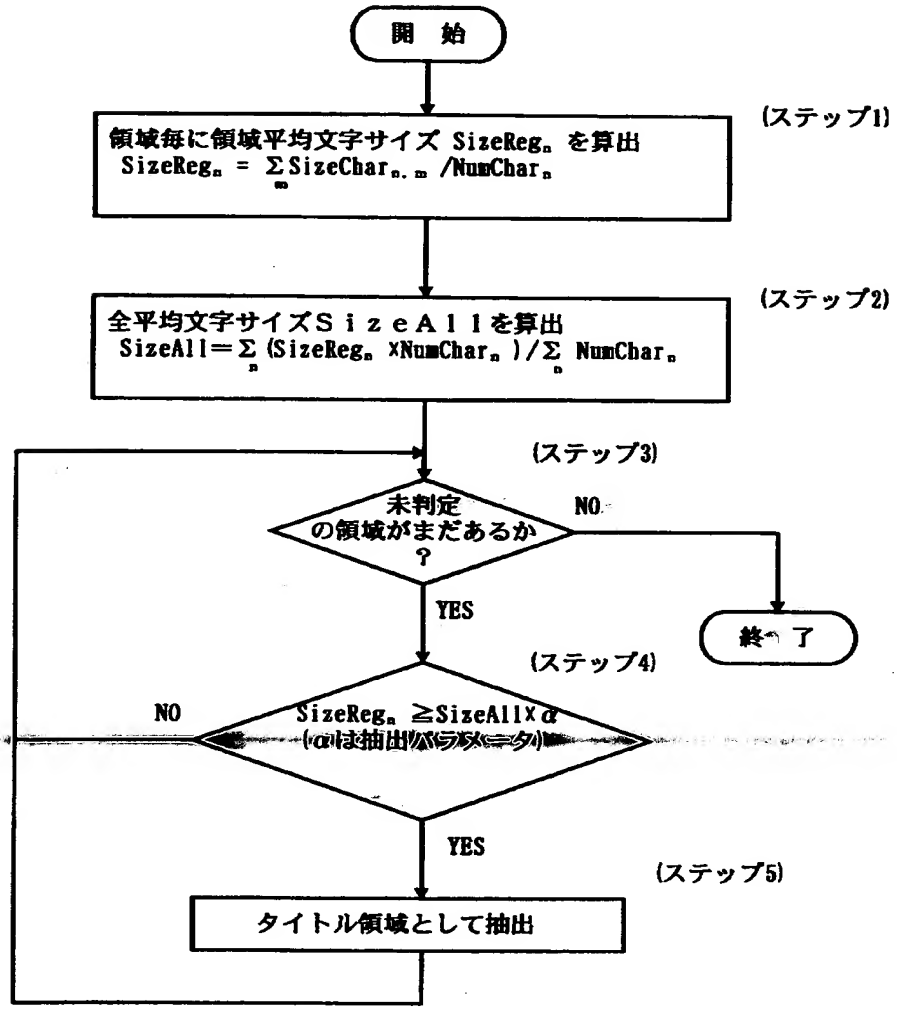
【書類名】　図面

【図１】


文字矩形
生成手段
2
文書画像
入力手段
1
領域分割
手段
3
タイトル領域
抽出手段
4
画像処理
手段
画像
メモリ
11b
7
文字認識
手段
5
c：タイトル領域
位置・サイズ
画像処理
手段
11a
タイトル領域
格納ポインタ
b：
文書登録
手段
6
a
b
c
8a
8b
8c
8 記憶手段
文書画像
タイトル領域
画像
登録情報
管理テーブル
Aa
Ab
Ac
a：
文書画像格納
ポインタ
表示制御 手段
10
指示入力手段
9

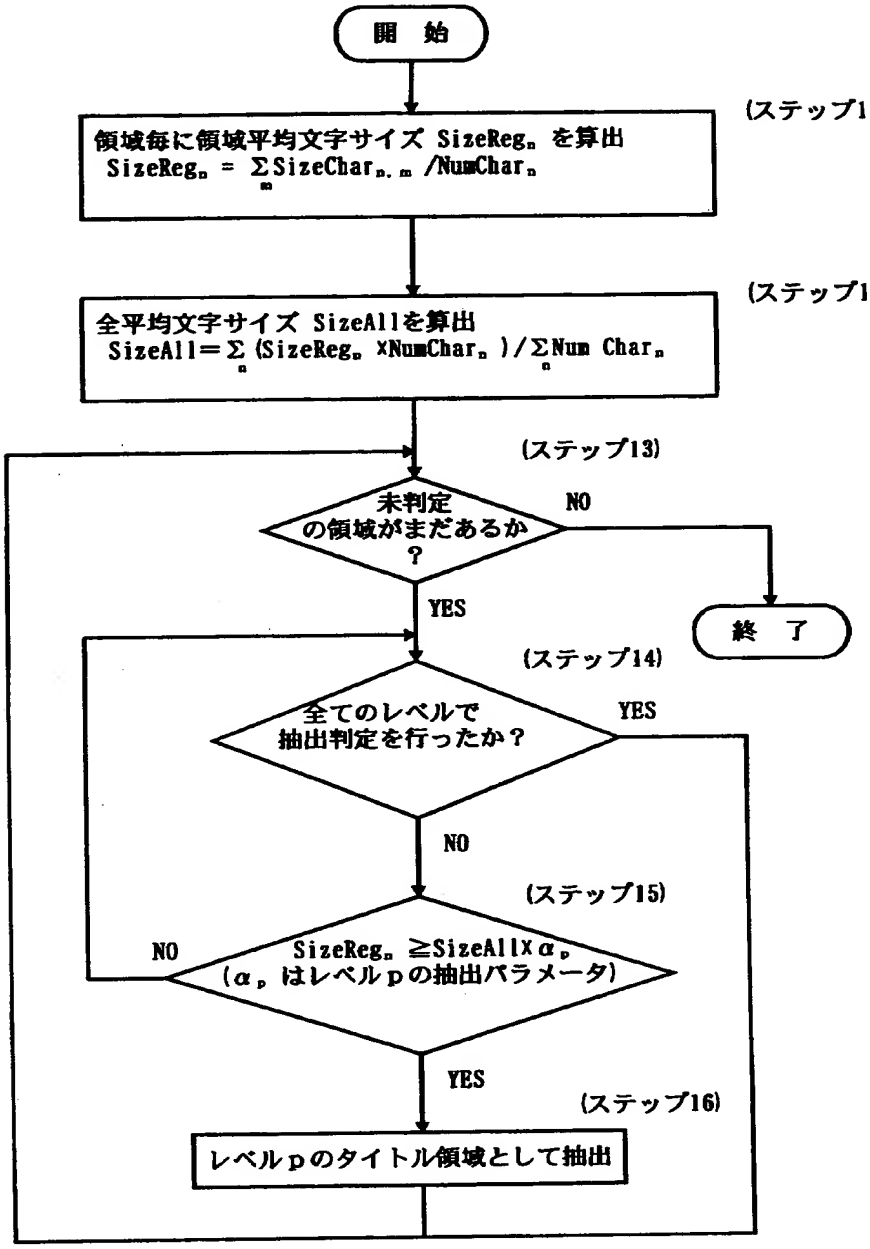
【図２】

第１の実施の形態におけるタイトル領域抽出処理のフローチャート



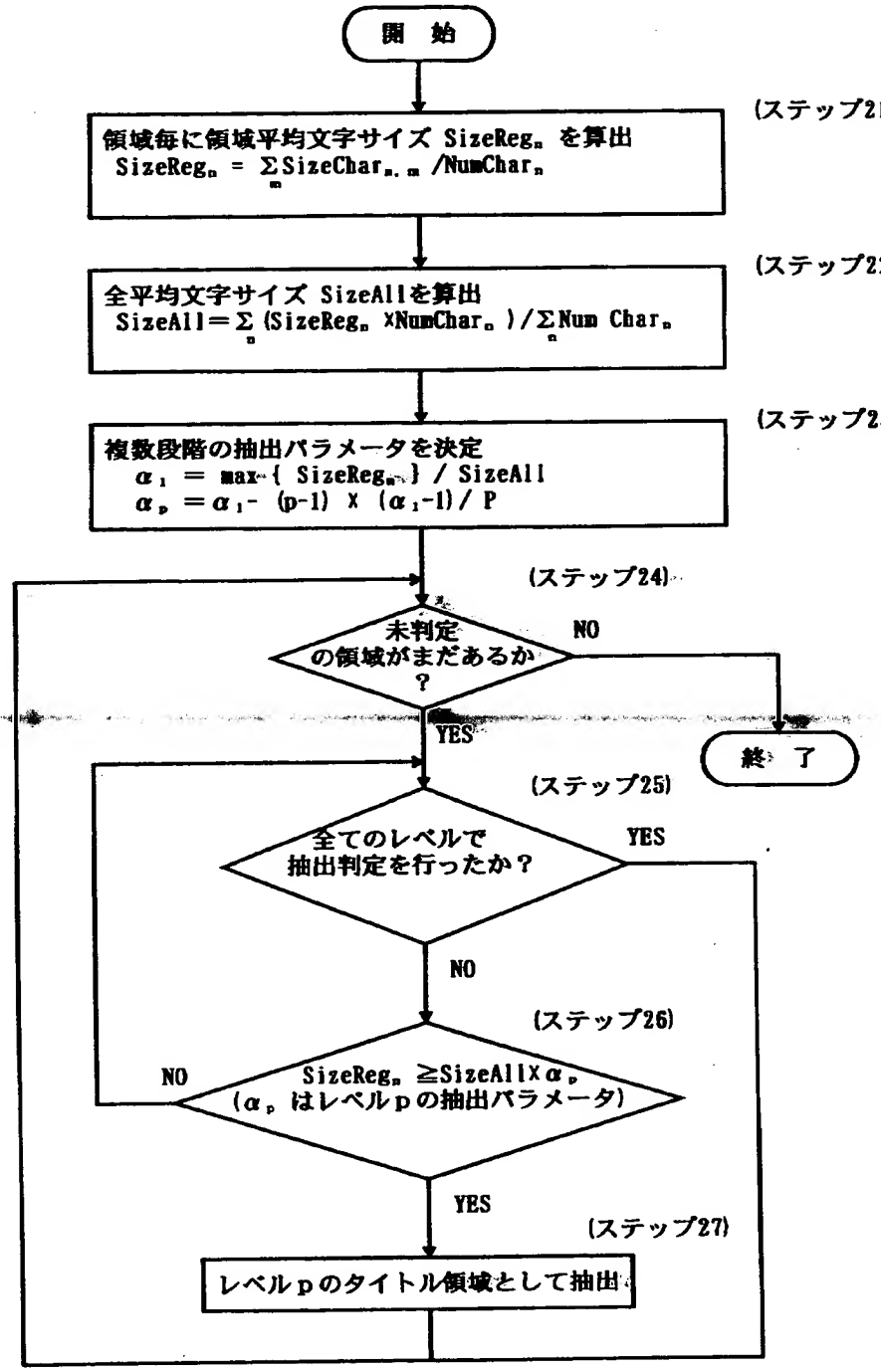
【図３】

第２の実施の形態におけるタイトル領域抽出処理のフローチャート



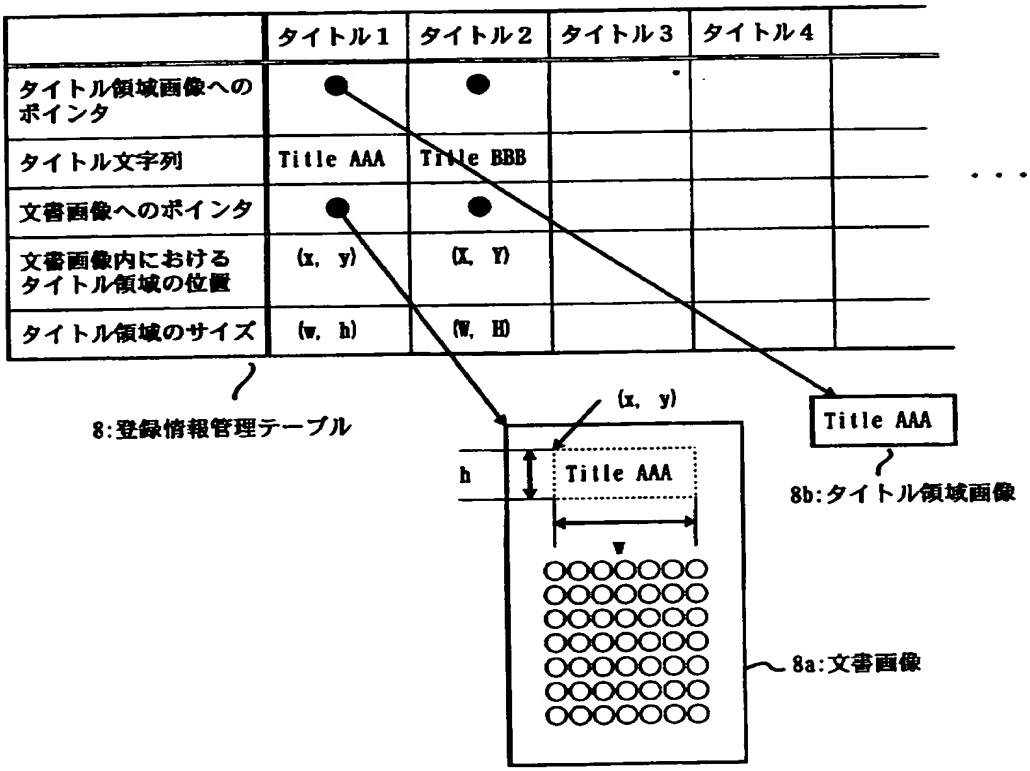
【図４】

第３の実施の形態におけるタイトル領域抽出処理のフローチャート



【図５】

第１の実施の形態における登録情報管理テーブルの説明図



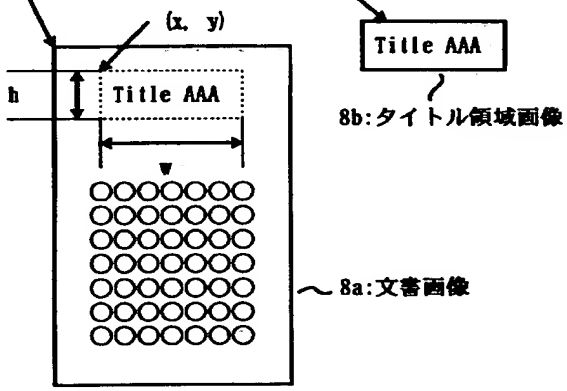
【図６】

第２の実施の形態における登録情報管理テーブルの説明図

	タイトル１	タイトル２	タイトル３	タイトル４	
タイトル領域画像へのポインタ	●	●			
タイトル文字列	Title AAA	Title BBB			
文書画像へのポインタ	●	●			
文書画像内におけるタイトル領域の位置	(x, y)	(X, Y)			
タイトル領域のサイズ	(w, h)	(W, H)			
レベル属性	1	3			
全レベル数	5	5			

・・・

8:登録情報管理テーブル

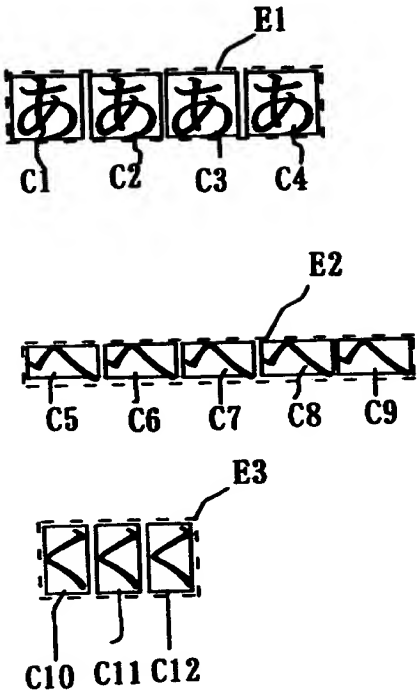


【図７】

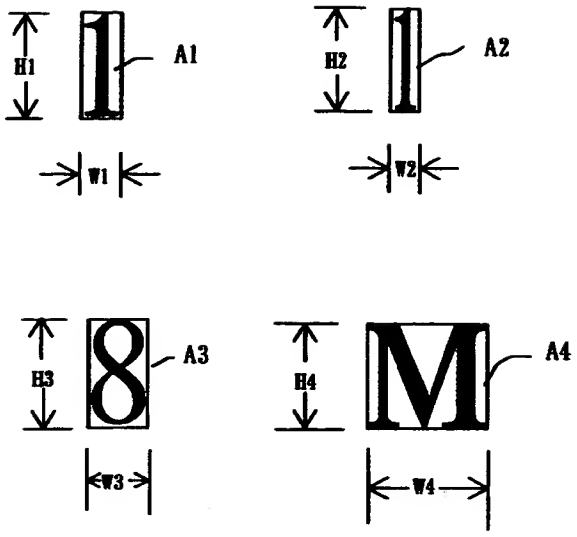
ラベリング処理の説明図

W１ (白)	W２ (黒)	W３ (黒)
W４ (白)	W５ (黒)	W６ (白)
W７ (白)	W８ (黒)	W９ (白)

【図８】



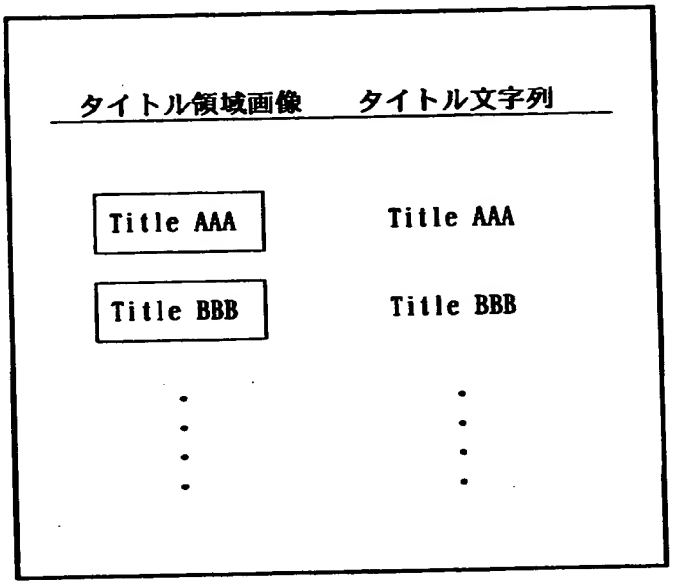
【図９】



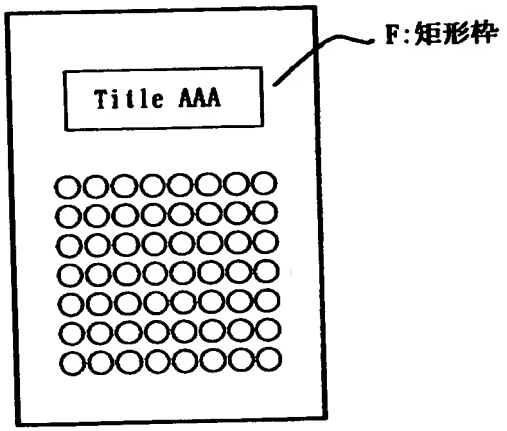
【図１０】

第１の実施の形態の検索時においてディスプレイに表示される内容

（Ⅰ）

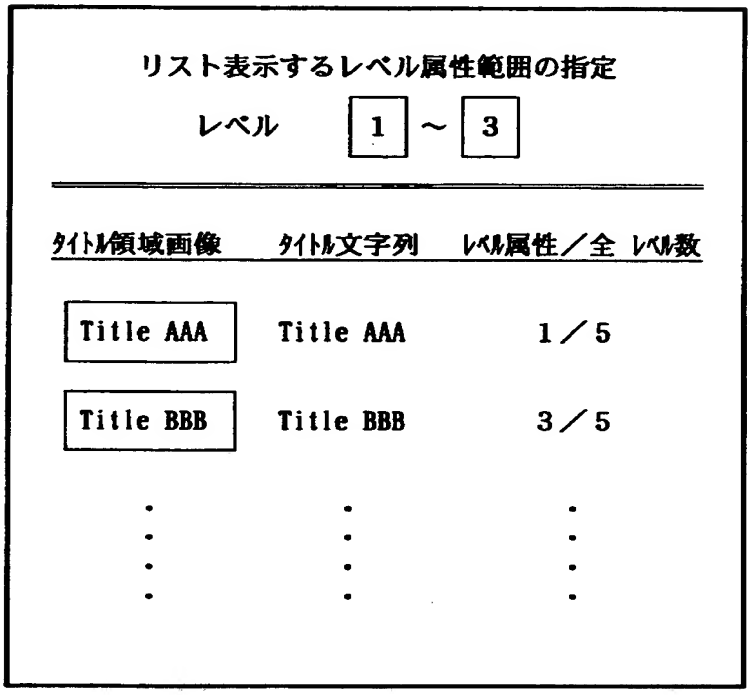


（Ⅱ）



【図１１】

第２の実施の形態の検索時においてディスプレイに表示される内容



【書類名】　要約書

【要約】

【課題】　文書画像よりタイトルを抽出する処理の精度を上げる。

【解決手段】　上記タイトル領域抽出手段４は、全領域における文字の平均高さに相当する全平均文字サイズを算出した後、該全平均文字サイズに抽出パラメータを乗算した抽出判定値と領域平均文字サイズとを比較し、上記抽出判定値より大きい領域平均文字サイズの領域をタイトル領域として抽出するようにしている。このようにすれば、上記抽出判定値より大きい領域平均文字サイズの領域であればタイトル領域として抽出されるため、１つの文書画像から複数のタイトル領域を抽出できることになる。

【選択図】　図　1

出証特平１１－３０９３６００

出　願　人　履　歴　情　報

識別番号　　〔０００００５８２１〕

１．変更年月日　　１９９０年　８月２８日
［変更理由］　新規登録
住　所　　大阪府門真市大字門真１００６番地
氏　名　　松下電器産業株式会社